State of the Art of Object Recognition Techniques

SCIENTIFIC SEMINAR

submitted by
Simon Achatz

NEUROSCIENTIFIC SYSTEM THEORY
Technische Universität München

Prof. Dr Jörg Conradt

Supervisor: Marcello Mulas, PhD
Final Submission: 06.07.2016

In your final hardback copy, replace this page with the signed exercise sheet.

Abstract

Object recognition plays an important role in computer vision. It is indispensable for many applications in the area of autonomous systems or industrial control.
In this work, several concepts of different object recognition approaches such as appearance-based and feature-based methods are presented. Furthermore, the function and pros and cons of state-of-the-art object recognition algorithms are described. SIFT and SURF as famous members of the feature-based strategy, convolutional neural networks and PCA and LDA, which belong to the appearance-based methods, will be introduced. Finally, these algorithms are compared in terms of, for example, complexity, accuracy and robustness. This work is concluded with an advise concerning the areas of application and the performance of each algorithm.

Contents

Chapter 1

Introduction

Computer vision is the ability of machines to see and understand what is in their surroundings. This field contains methods for acquiring, processing and analysing of images to be able to extract important information used by artificial systems. Most recently, in computer vision a lot of research is conducted, especially in its major sub-domains such as object recognition, motion analysis or scene reconstruction.
Object recognition describes the task of processing an image in a certain way to localize and to classify objects. There exists a huge variety of object recognition approaches, but the general concept remains the same. An object recognition system uses training datasets containing images with known and labelled objects and it extracts different types of information based on the chosen algorithm. This can be information about colours, edges, geometric forms and so on. Generally, for any new image the same information is gathered and compared to the training dataset to find the most suitable classification.
Many applications using object recognition can be found in everyday life. Starting with robots in industrial environments, face or handwriting recognition and up to autonomous systems such as modern cars which use object recognition for pedestrian detection, emergency brake assistant and so on.
At first, in Chapter 2, a short overview of different strategies for object recognition is stated. Common approaches such as appearance-based and feature-based methods are mentioned as well as little known methods such as interpretation trees or template matching and the rising star: artificial neural networks. Followed by the description of five selected, popular object recognition algorithms in Chapter 3. The functionality of PCA and LDA as examples for appearance-based methods, SIFT and SURF for feature-based methods and convolutional neural networks are presented. In Chapter 4, the selected algorithms are analysed and compared in terms of robustness, accuracy and accuracy. Finally, the conclusion sums up the result and states which algorithm outperforms the others or is useful in which case of application.

Chapter 2

Characterization of General Object Recognition Strategies

In this chapter, general strategies on which the most popular object recognition algorithms are based on, are characterized. A lot of algorithms and combinations of them are based on the first two methods: appearance-based and feature-based. Then, two more methods, interpretation trees and pattern matching are introduced and last but not least, the concept behind artificial neural networks is described.

2.1 Appearance-based Method

First, a short overview of the concept of the appearance-based object recognition strategy is presented. Appearance-based methods are popular for face or handwriting recognition. For this strategy, a set of reference training images, which are highly correlated, is needed. For example, 100 images of faces and a set of images containing background or random objects. This dataset is compressed using dimensionality reduction techniques to obtain a lower dimension subspace, also called eigenspace. Parts of the new input images are projected on the eigenspace and then correspondence is examined. [HLC12][MK01]
More details about appearance-based methods in Chapter 3, when describing the two famous algorithms PCA and LDA.

2.2 Feature-based Method

The next strategy is called feature-based, because algorithms recognise objects based on specific features. Features are supposed to be characteristic for each object, often one object is not only described by one attribute but multiple features. Colours, contour lines, geometric forms or edges (gradient of pixel intensities) are popular choices. The basic concept of feature-based object recognition strategies is following:

Every input image is searched for a specific type of feature, this feature is then compared to a database containing models of the objects in order to verify if there are recognised objects. [AT13, MO04]

As already mentioned, object features can have many faces, but simplified spoken, they all can be divided in just two categories. Features and their descriptors can be either found considering the whole image (global feature) or after observing just small parts of the image (local feature).

An histogram of the pixel intensity or colour are simple examples for global features. It is not always reasonable to compare the whole image, as already slight changes in illumination, position (occlusion) or rotation lead to significant differences and a correct recognition is not possible anymore. [GL]

Descriptors of local features are more robust against these problems and therefore algorithms with local features often outperform global feature-based methods. In Fig. 2.1, the general concept of local feature-based algorithms can be seen. Two small patches are compared and not the whole image, these patches may be rotated and normalized first in order to achieve higher accordance. This approach has lead to much progress on research in object recognition. In Chapter 3, the local feature strategy is further described with referencing to two famous algorithms: SIFT and SURF.

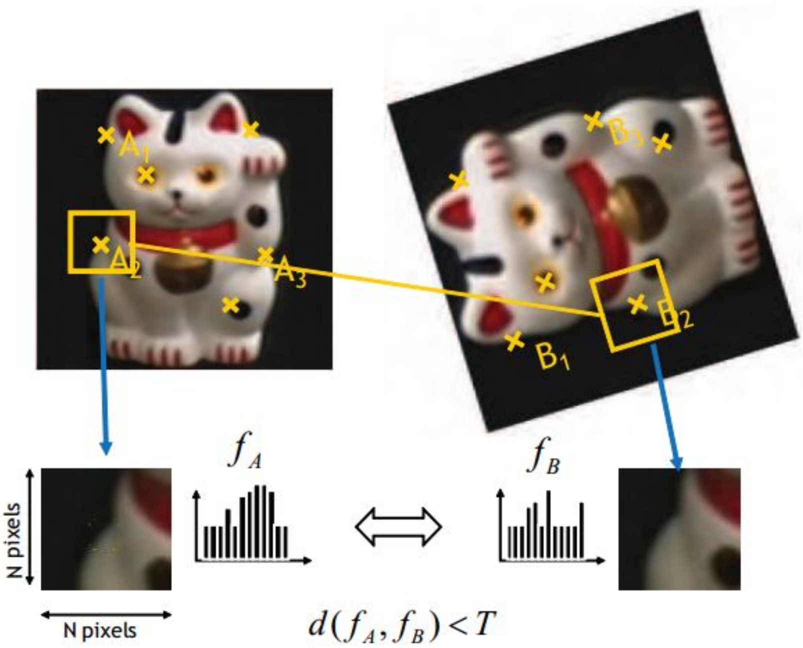

Figure 2.1: Two patches of different images are cut and compared if the error between the patches is below a certain threshold.

2.3 Interpretation Tree

The next object recognition strategy is called interpretation tree which is a depth-first search algorithm for model matching. Algorithms based on this approach often try to recognise n-dimensional geometric objects, therefore a database containing models with known features is necessary. The feature set might consist of distance, angle and direction constraints between points on the surface of the objects. [AT13, Hag03]
In fig. 2.2, an example of an interpretation tree is shown. In the first level, a feature of the model is chosen and then compared to all features from the unrecognised object. This procedure is repeated until all combinatorial options are gone through.

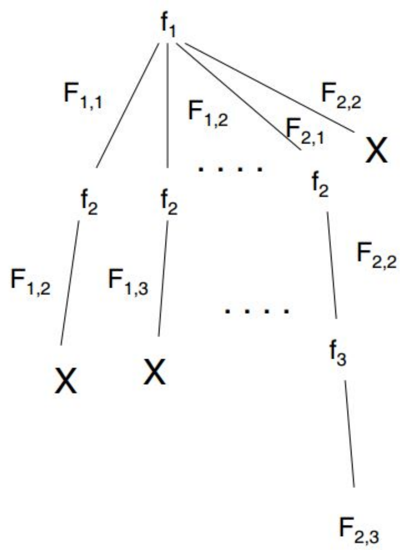

Figure 2.2: Procedure of an interpretation tree algorithm.

But due to the unpopularity of this strategy in object recognition, it will not be further considered in this work.

2.4 Pattern Matching

Methods of pattern matching, or sometimes called template matching, are often used because of their simplicity. Template matching is a technique for finding small parts of an image which match a template image. [EASD00, KA13]
The computation is quite easy: In Eq. (2.1), the squared differences between an image patch I and a template M are summed pixelwise. A threshold has to be provided in order to let the algorithm decide whether a template matched and an

object was recognised.

$$r(x,y) = \sum_{i \epsilon M} \sum_{j \epsilon M} (I(x+i, y+j) - M(i,j))^2 \tag{2.1}$$

The result can be adjusted in order to stabilize against small distortion and light changes with Eq. (2.2), where n is the number of pixels in the template.

$$r = \frac{\sum IM - \sum I * \sum M}{\sqrt{(n \sum I^2 - (\sum I)^2)(n \sum M^2 - (\sum M)^2)}} \tag{2.2}$$

One famous application of template matching is traffic sign recognition, small parts of the input image are tried to be matched with a database full of different images of traffic signs. As this approach has lots of disadvantages such as problems with occlusion, rotation, scaling, different illuminations and so on, it will not be given further attention in this work.

2.5 Artificial Neural Networks

Artificial neural networks are models inspired by biological neural networks. Such a model consists of several layer, as it can be seen in Fig. 2.3, in which each layer is composed of a certain number of neurons. An input and an output layer is the minimum amount of layers a network can have, but normally hidden layer are included to be able to learn more complex things such as object recognition.

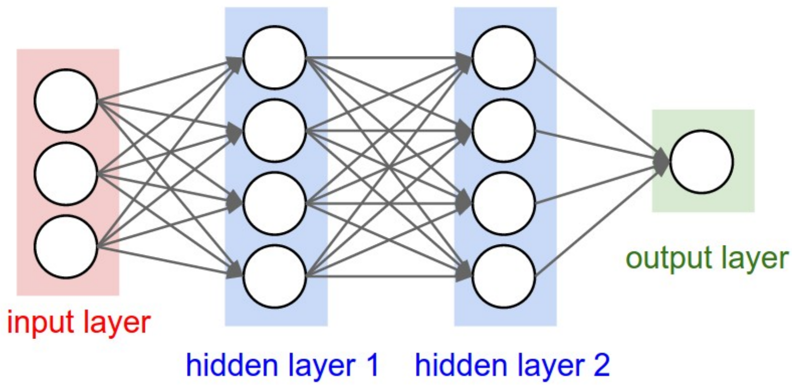

Figure 2.3: A neural network containing one input layer, two hidden layer and one output layer.

All neurons from one layer are connected to all neurons from the next layer and therefore create a huge network with millions of parameters. All of these connections have a weight which is updated during learning phase. Neurons are activated if the sum of the input signals is above a certain threshold and a activation function triggers the output. (see Fig. 2.4)

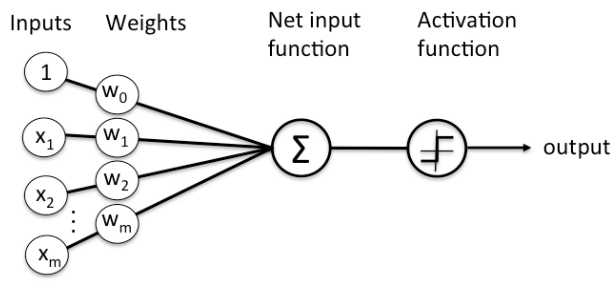

Figure 2.4: A neural network containing one input layer, two hidden layer and one output layer.

There are different types of networks such as feed-forward, recurrent with different number and types of hidden layers, while the input (e.g. number of pixels) and output (number of classes) layer are fixed. Later, convolutional neural networks and their hidden layers are explained in a more detailed way in Chapter 3. New inputs go through the same way, some neurons might be activated based on the trained network and finally, this leads to the most suitable classification. [KR14, STE13]

Chapter 3

Review of Popular Object Recognition Algorithms

Previously, general object recognition strategies were presented. In this chapter, famous algorithms from the most promising approaches are demonstrated. Short descriptions and the general function of SIFT and SURF, which are examples for the feature-based approach, PCA and LDA, which are appearance-based methods, and convolutional neural networks are displayed.

3.1 SIFT - Scale-Invariant Feature Transform

Lowe introduced his very well known feature-based scale-invariant feature transform (SIFT) algorithm in 2004. There exists a lot of variations of the algorithms and combinations of SIFT with other approaches in order to wipe out disadvantages. They all consist of four major stages (see [Low04]), but in this work, it is focused on the general SIFT:

- Scale-space extrema detection
- Keypoint localization
- Orientation assignment
- Keypoint descriptor

In the first stage, extrema are searched over all image locations and all scales. This can be achieved easily with the difference-of-Gaussian (DOG) function, which is explained in Fig. 3.1. The left side contains the results of a convolution of the input image I with a Gaussian filter G, which is computed for all scales. The difference of two adjacent scales is the DOG.
In the next two steps, the extrema are localized and the orientation of each maximum is computed. Maxima and minima of the DOG images are found by comparing a

pixel with its 26 neighbours. 8 neighbours in the current scale and 2*9 neighbours in the adjacent scales. [RW08, Zho04]

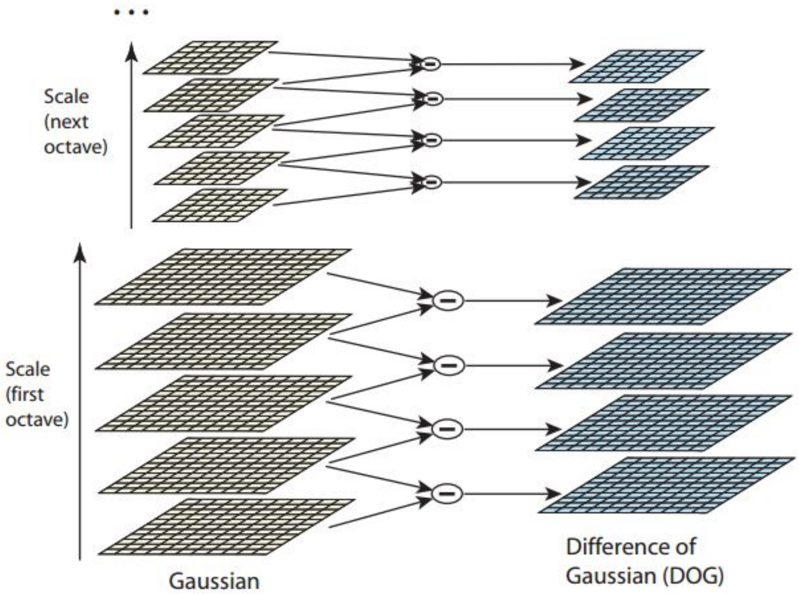

Figure 3.1: Different scales of the Gaussian filter, which are subtracted to obtain the difference-of-Gaussian

Finally, keypoint or feature descriptors are introduced in the last stage. Each local feature generates a descriptor which has information about the gradient magnitude and orientation around the sample point. That region is weighted with a Gaussian window and 4x4 sub-regions are combined into a orientation histogram. (see Fig. 3.2)

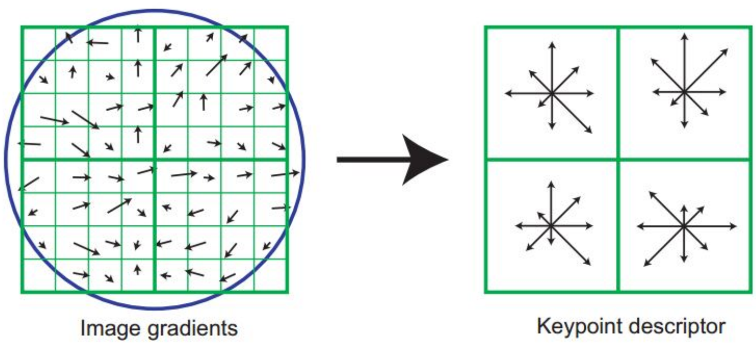

Figure 3.2: 64 image gradients (magnitude and orientation) are combined to 2x2 descriptors containing histograms of the orientation.

3.2 SURF - Speeded-Up Robust Features

Bay et al. used SIFT as a model for their more advanced speeded-up robust feature (SURF) algorithm. This feature-based algorithm consists of the same steps as SIFT, but introduces significant innovations.
In the beginning, features have to be localized first. SURF uses integral images and the Hessian matrix in order to detect maxima. The Hessian matrix consists of approximations of the convolution of Gaussian second order derivatives and the image. [BETG08]

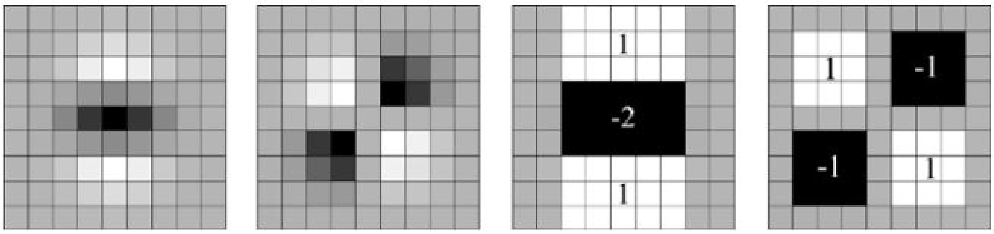

Figure 3.3: Left side: Gaussian second order derivatives in y- and xy-direction; Ride side: approximations of left side.

The second part of the algorithm is the description of interest points. First, the orientation of each square of a 4x4 grid each cell containing 4 samples around a feature is calculated using Haar wavelet responses.

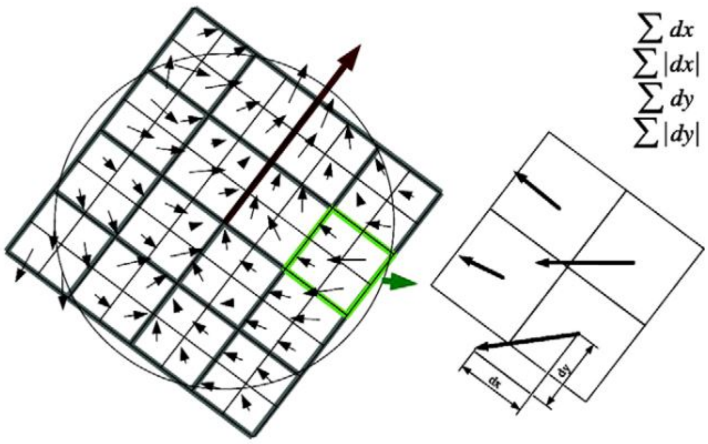

Figure 3.4: 4x4 square grid over interest point with each cell containing 2x2 samples.

The 4D feature descriptor vector v is constructed out of the sums of the responses d over each sub-region:

$$v = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|) \tag{3.1}$$

3.3 PCA - Principle Component Analysis

Principle component analysis (PCA) is a procedure of statistics which reduces the dimensionality of a collection of observed data. In general, PCA is a orthogonal linear transformation which transforms data to a new coordinate system. The first coordinate equals to the direction of the greatest variance of the data, this is called first principle components. The second principle component equals to the second greatest variance and so on.

In Fig. 3.5, the two axis λ_1 and λ_2 represent the two principle components of the observed data, where λ_2 is the first principle component.

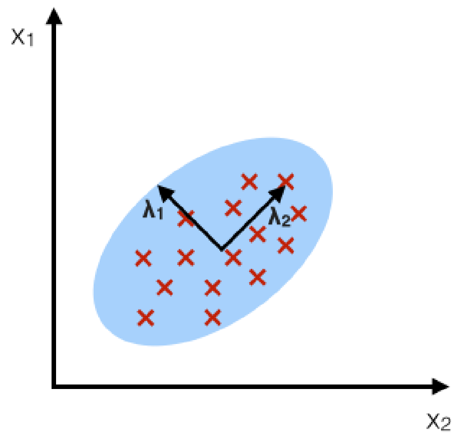

Figure 3.5: Collection of data with two principle components λ_1 and λ_2.

In PCA for object recognition mostly the first principle component is needed as it still contains most information about the data. The data is projected on the first principle component in order to maintain maximum variance.

$$w_{(1)} = \underset{||w||=1}{argmax} \left\{ ||Xw||^2 \right\} = argmax \left\{ \frac{w^T X^T X w}{w^T w} \right\} \tag{3.2}$$

The weights of the first principle component can be calculated using Eq. 3.2, which is in matrix format, where X is a $m * n$ matrix containing m n-dimensional samples.

PCA is a popular technique for pattern and object recognition, however, it is not suitable for classification as separation does not work well. In the next section, an alternative method is proposed.

3.4 LDA - Linear Discriminant Analysis

Another famous appearance-based algorithm is called linear discriminant analysis (LDA). It is a method used in statistics for dimensionality reduction or classification. Each class is represented by mean μ_i and the same covariance Σ. The algorithm minimizes the intra-class variance Σ, while the inter-class variance Σ_b is maximized. With C as the number of classes and μ as the mean of the class means one obtains:

$$\Sigma_b = \frac{1}{C}\sum_{i=1}^{C}(\mu_i - \mu)(\mu_i - \mu)^T \tag{3.3}$$

In Eq. 3.4, the class separation S in direction $\vec{w}$ is calculated:

$$S = \frac{\vec{w}^T \Sigma_b \vec{w}}{\vec{w}^T \Sigma \vec{w}} \tag{3.4}$$

A projection is good, if it separates classes well like it is shown in Fig. 3.6. The projection on the y-axis yields a bad projection as the two classes are not separable anymore, while a projection in x-direction provides a good result.

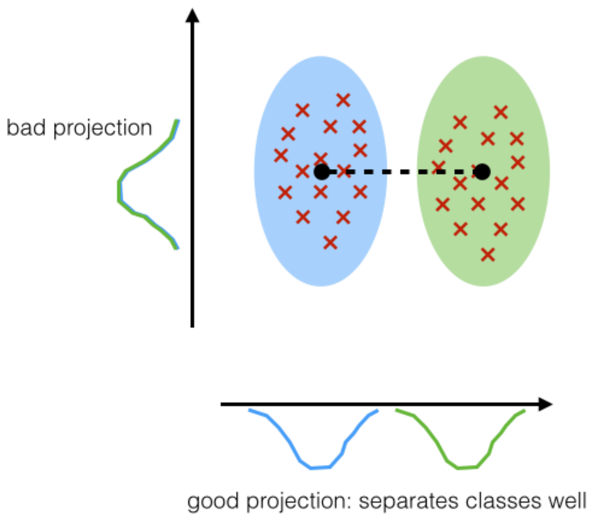

Figure 3.6: A projection can lead to good or bad class separation.

As means and covariance are not always known, different approaches have to be applied to still be able to use LDA. Maximum likelihood estimation or maximum a posteriori estimation can help.

3.5 CNN - Convolutional Neural Network

In Chapter 2, the general concept behind artificial neural networks was explained. There is one special type of neural networks which is particularly suitable for object recognition: (deep) convolutional neural networks (CNN). The term *deep* means that there is at least one hidden layer and *convolutional* implies the use of convolution layers. The basic principles of CNNs are inspired by the biological visual cortex of humans.
The architecture of an example CNN can be seen in Fig. 3.7. Input images with 28x28 pixels are convoluted with a filter to obtain 3D feature maps. The succeeding sub-sampling, or often called pooling, layer further reduces the amount of data. This procedure is continued until a one-dimensional vector, which represents the different classes, is obtained.

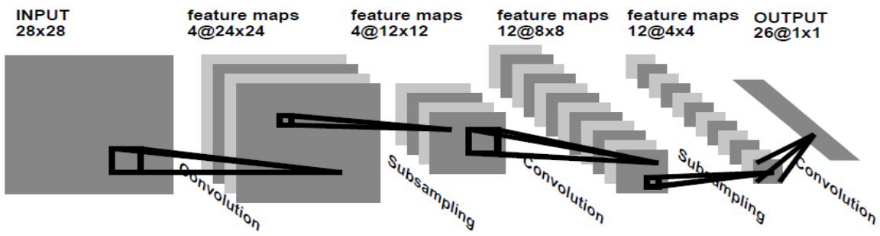

Figure 3.7: One example architecture of a convolutional neural network using sub-sampling and convolution hidden layers.

As most of the object recognition algorithms, CNNs need a training to adapt all weights of the neurons. During that phase, different levels of features are extracted (see Fig. 3.8). Low-level features contain colour, lines or contrast, whereas edges and corner belong to mid-level features. High-level features already include class specific forms or sections. [Jos15]

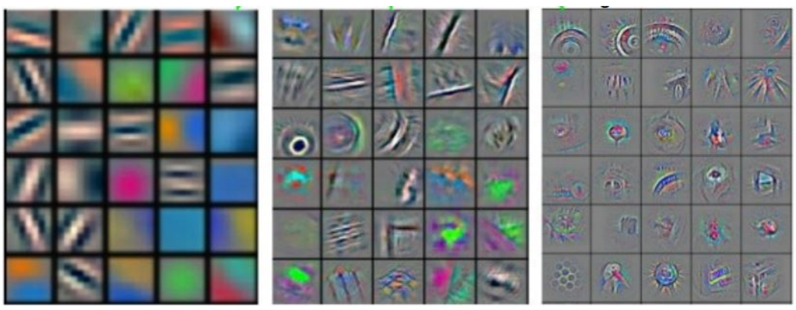

Figure 3.8: Intermediate results from hidden layers. From left to right: low-level, mid-level and high-level features.

After the introduction of five popular object recognition algorithms, it is about time to compare them and to see which algorithm is useful for which application.

Chapter 4

Performance Analysis

In the preceding chapter, five algorithms from different object recognition strategies were explained. First, these algorithms are compared regarding invariances and robustness, that includes weaknesses in certain situations such as occlusion or rotation. Computational load or memory usage are points which are examined in the comparison about complexity. The last point contains results in reliability and accuracy, for which object recognition challenges are introduced.

4.1 Invariances and Robustness

First, the algorithms are analysed and checked whether invariances occur and what level of robustness they have. It is easy to recognise objects when circumstances such as light condition, position and rotation of the object remain the same during training and testing. However, this is of course not always the case and is not wanted. The object is supposed to be recognised also in different environmental situation such as at night or with fog or rain.
In [PBCD11, JG09, Jos15] following findings were won:

Table 4.1: Performance evaluation in terms of robustness

	SIFT	SURF	PCA	LDA	CNN
rotation	++	0	+	+	++
illumination	0	++	-	-	++
occlusion	++	+	−	−	++
speed	+	++	0	0	-

CNNs just depend on the variations of the images in the training dataset, but its complex architecture provides very good results even when parts of the object are occluded or even human eyes cannot recognise the object. Only disadvantage is the comparable low speed.

SURF outperforms SIFT, because it uses approximations to be faster and adjusts parts of the steps to wipe out the problems of SIFT. In comparison with the other algorithms, SIFT and SURF are faster, because only small parts of the image or features have to be analysed, instead of CNNs using the whole image.

PCA and LDA have their problem with occlusion. The object to be detected has to be fully visible and has to be easily separated from the background, because the algorithms depend on the complete appearance of the object.

4.2 Complexity

Secondly, the algorithms are compared regarding complexity, especially in terms of computational load and memory usage.

The computational load and the memory usage highly depends on the number of images used in the training dataset. PCA, for example, is popular for face recognition, therefore for this application just training images of faces and non-faces are needed. Whereas CNNs are often used to identify many more different objects. In order to achieve good results with CNNs all objects have to be represented in the training dataset.

This fact is more or less independent of the chosen algorithm, but there are also significant differences during runtime after the learning phase.

The appearance-based algorithms rely on relatively easy matrix equations, that is why the memory usage and computational load is rather low. PCA and LDA do not need to save models to compare features, it is enough to have one representation for object 1, object 2 and so on. Regarding complexity, SURF outperforms SIFT (see [JG09]) as it mostly uses approximations of the steps of SIFT. In comparison to the appearance-based methods, SIFT and SURF need a database with models to match the features, which results in a higher memory usage. CNNs are very powerful, but need a lot of resources as up to one billion parameters (weights of neurons) were utilized.

In the following Tab. 4.2, the gained knowledge is summarized and the results are rated from -- to ++.

Table 4.2: Performance in terms of complexity

	SIFT	SURF	PCA	LDA	CNN
computational load	+	+	+	0	--
memory usage	-	0	+	+	-
overall complexity	0	+	++	+	--

4.3 Reliability and Accuracy

For training and testing of new object recognition approaches and algorithms, huge datasets containing labelled images with known objects were created. ImageNet, for example, consists of about 80 million images with over 80000 different objects and 1000 images each. Pascal VOC and AlexNet are further examples. There is an associated challenge in each dataset included, in which developers can rank their algorithms.
A couple of years ago, the best object recognition algorithms achieved accuracy rates of about 70 %. Within the last few years and the introduction of deep learning strategies, new dimensions were reached. Very recent and complex algorithms achieve results far beyond the 90 % (see Fig 4.1).

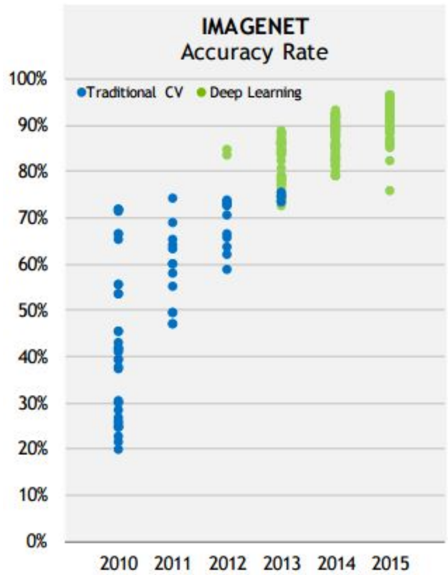

Figure 4.1: The development of accuracy rates of traditional computer vision and deep learning regarding ImageNet. [NVI16]

It is hard to find reliable literature which compares all the algorithms mentioned in this work. Nonetheless, it is proven (see [PBCD11, JG09]), that SIFT results in a better object detection than SURF due to the fact that SURF uses approximations to save time and effort.
Regarding 3.3 and 3.4, one might suggest, that LDA is superior than PCA as it uses class discrimination and is more suitable for classification. In [MK01], Martinez and Kak prove, that in cases where the number of samples in the training dataset is small PCA outperforms LDA.
When relatively low speed and high computational load is not a problem, then CNNs definitely are the best choice for object recognition.

Chapter 5

Conclusion

Five different concepts of object recognition approaches were presented in Chapter 2. Interpretation trees and template matching were examples for very specific and easy strategies and therefore not further given special consideration. Feature-based and appearance-based methods as well as artificial neural networks were treated and assessed as good object recognition approaches. After introducing the general concepts, some algorithms were explained in a more detailed way in Chapter 3. First, SIFT and SURF, which act as local feature-based algorithms, were reviewed. Two appearance-based algorithms (PCA and LDA) were explained afterwards. Regarding deep neural networks, CNNs were chosen as a special type of network which is popular in image processing and object recognition. As a last step, these five algorithms were analysed and evaluated in terms of accuracy, complexity and robustness.

The main strength of SIFT is its comparable good recognition rate, whereas its processing speed is low. However, SURF outperforms the other algorithms in terms of speed and complexity, but it lacks in performance regarding recognition rate.

PCA also has its right to exist in object recognition. In general, it does not achieve a good accuracy, but in some special cases it outperforms other algorithms. The main advantage of PCA is the fact that it reduces dimensionality and is therefore often used in combination with other approaches. LDA, the second appearance-based method, has good results in memory usage and computational load and is very popular in two-class applications.

Convolutional neural networks achieve the best results so far. Using complex architectures, it is possible to reach accuracy rates of about 95 %. Despite this impressing outcome, CNNs cannot manage without negative impacts. Very huge training datasets and up to billion parameters lead to a high computation load and memory usage, which then needs high processing power to be able to be applied usefully.

It can be seen, that every object algorithm has different advantages and disadvantages. Hence, it is almost not possible to create a complete, meaningful ranking, as too many different aspects have to be considered. It depends on the target application which algorithm shall be used.

A face recognition system could be based on a lightweight PCA or LDA and it is not worth to create a complex and powerful multi-layer neural network. Once developing autonomous systems, for example self-driving cars, it is advisable to use the best available algorithm, as it may lead to fatal accidents. CNNs are the most popular choice of car manufactures to monitor the surroundings of the car. In this case, it is acceptable to have a learning duration of the neural network of several days.
Although the concept of neural networks is way older than the other approaches, it still can compete with them very well. Especially the progress in the development of better and faster hardware, it is no problem anymore to cope with the vast amount of parameters and equations. With regard to the ongoing and continuous enhancement of CPUs and GPUs, deep neural networks and in particular CNNs will be most likely the common way for object recognition in future.

List of Figures

Bibliography

[AT13] Alexander Andreopoulos and John K Tsotsos. 50 years of object recognition: Directions forward. *Computer Vision and Image Understanding*, 2013.

[BETG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). In *Computer Vision and Image Understanding 110*, 2008.

[EASD00] Markus Ehrenmann, Despina Ambela, Peter Steinhaus, and Rudiger Dillmann. A comparison of four fast vision based object recognition methods for programing by demonstration applications. In *IEEE International Conference on Robotics & Automation*, 2000.

[GL] Kristen Grauman and Bastian Leibe. Visual object recognition. Synthesis Lectures in Computer Vision.

[Hag03] G.D. Hager. Object recognition techniques. http://www.cs.jhu.edu/ hager, 2003.

[HLC12] Gee-Sern Hsu, Truong Tan Loc, and Sheng-Lun Chung. A comparison study on appearance-based object recognition. In *International Conference on Pattern Recognition*, 2012.

[JG09] Luo Juan and Oubong Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP)*, 2009.

[Jos15] Josip Josifovski. Object recognition: Sift vs convolutional neural networks. University Hamburg - Lecture Intelligent Robotics WS 2015/16, 2015.

[KA13] Khushboo Khurana and Reetu Awasthi. Techniques for object recognition in images and multi-object detection. *International Journal of Advanced Research in Computer Engineering & Technology*, 2013.

[KR14] Seyed-Mahdi Khaligh-Razavi. What you need to know about the state-of-the-art computational models of object-vision: A tour through the models. University Cambridge, 2014.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. University of British Columbia, 2004.

[MK01] Aleix M. Martinez and Avinash C. Kak. Pca versus lda. In *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2001.

[MO04] Jiri Matas and Stepan Obdrzalek. Object recognition methods based on transformation covariant features. In *12th European Signal Processing Conference*, 2004.

[NVI16] NVIDIA. Deep learning on gpus. http://on-demand.gputechconf.com/gtc/2015/webinar/deep-learning-course/intro-to-deep-learning.pdf, mar 2016.

[PBCD11] Nicolas Pinto, Youssef Barhomi, David D. Cox, and James J. DiCarlo. Comparing state-of-the-art visual features on invariant object recognition tasks. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 2011.

[RW08] Peter M. Roth and Martin Winter. Survey of appearance-based methods for object recognition. Technical report, Institute for Computer Graphics and Vision, Graz University of Technology, Austria, 2008.

[STE13] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems 26*, 2013.

[Zho04] Qiang Zhou. *Generalized Landmark Recognition in Robot Navigation.* PhD thesis, Ohio University, 2004.

License